A supervised learning for chromosome assignment for genetic markers without a reference genome

Emi Tanaka J. Taylor B. Cullis

School of Mathematics and Applied Statistics
University of Wollongong

Motivation

- A quantitative trait locus (QTL) is a section of the DNA (locus) that is linked to, or contains, the genes that control the quantitative trait.
- QTL analysis is often an important early step for identification of genes that cause trait variation.
- Often in crops, bi-parental population and genetic markers such as SNP, DArT, SSR are used to detect potential QTLs.
- A linear mixed model approach to QTL analysis accommodates well to account for non-genetic sources of variation and this is the basis of our QTL analysis.

Motivation

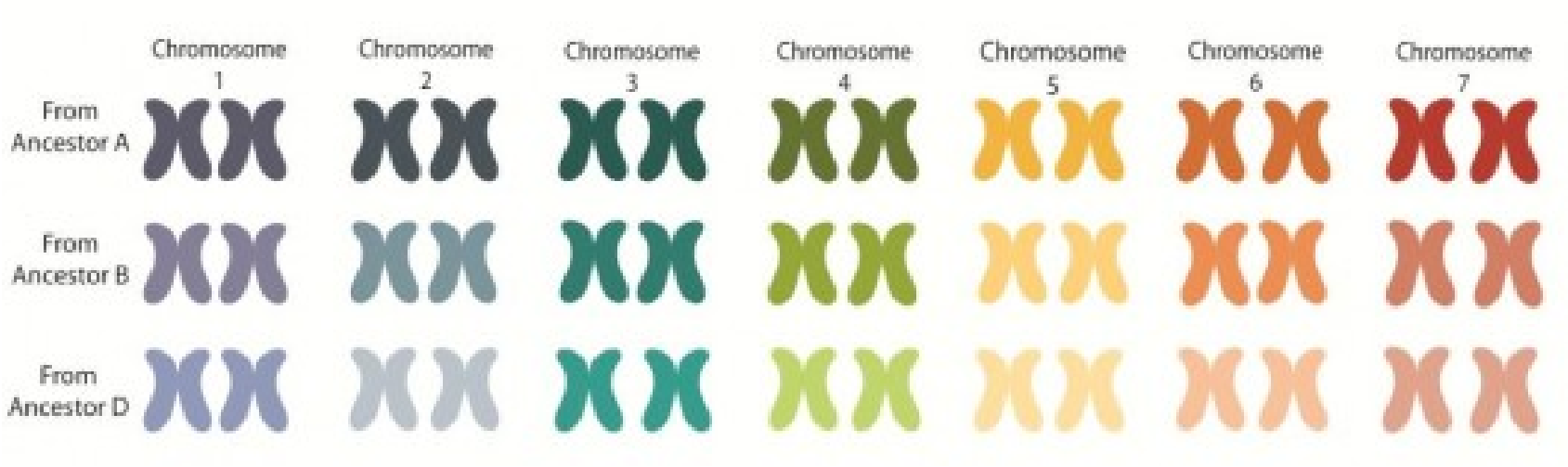
- Verybyla et al. (2007) proposed a mixed model approach that considered all markers/intervals (unlinked to QTL) as random effects.
- However these marker effects were considered iid.
- A number of different spatial covariance structure to the marker effects within chromosomes were proposed [Gianola et al., 2003, Smith and Cullis, 2011, Yang and Tempelman, 2012, Morota et al., 2014].
- These methods require a distance metric between markers and/or some ordering of markers.

Aim

We present a possible ordering and distance metric to be used for correlated marker effects in the QTL analysis.

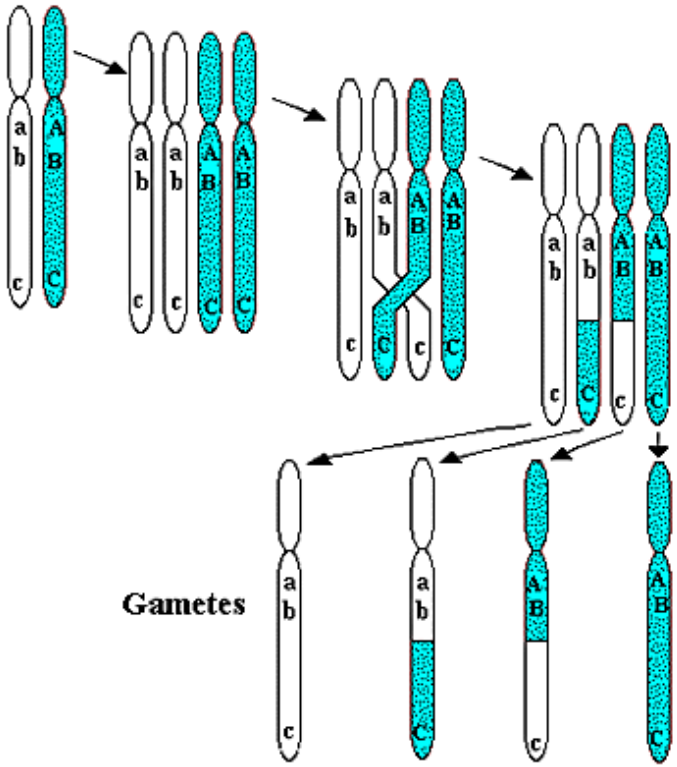
- A number of biological properties/assumptions are considered to build an appropriate ordering and chromosome assignment.

Biology Background



- Wheat is a hexaploid that has six copies of its seven chromosomes.
- We can treat wheat genome as 21 pairs of chromosomes.

Recombination



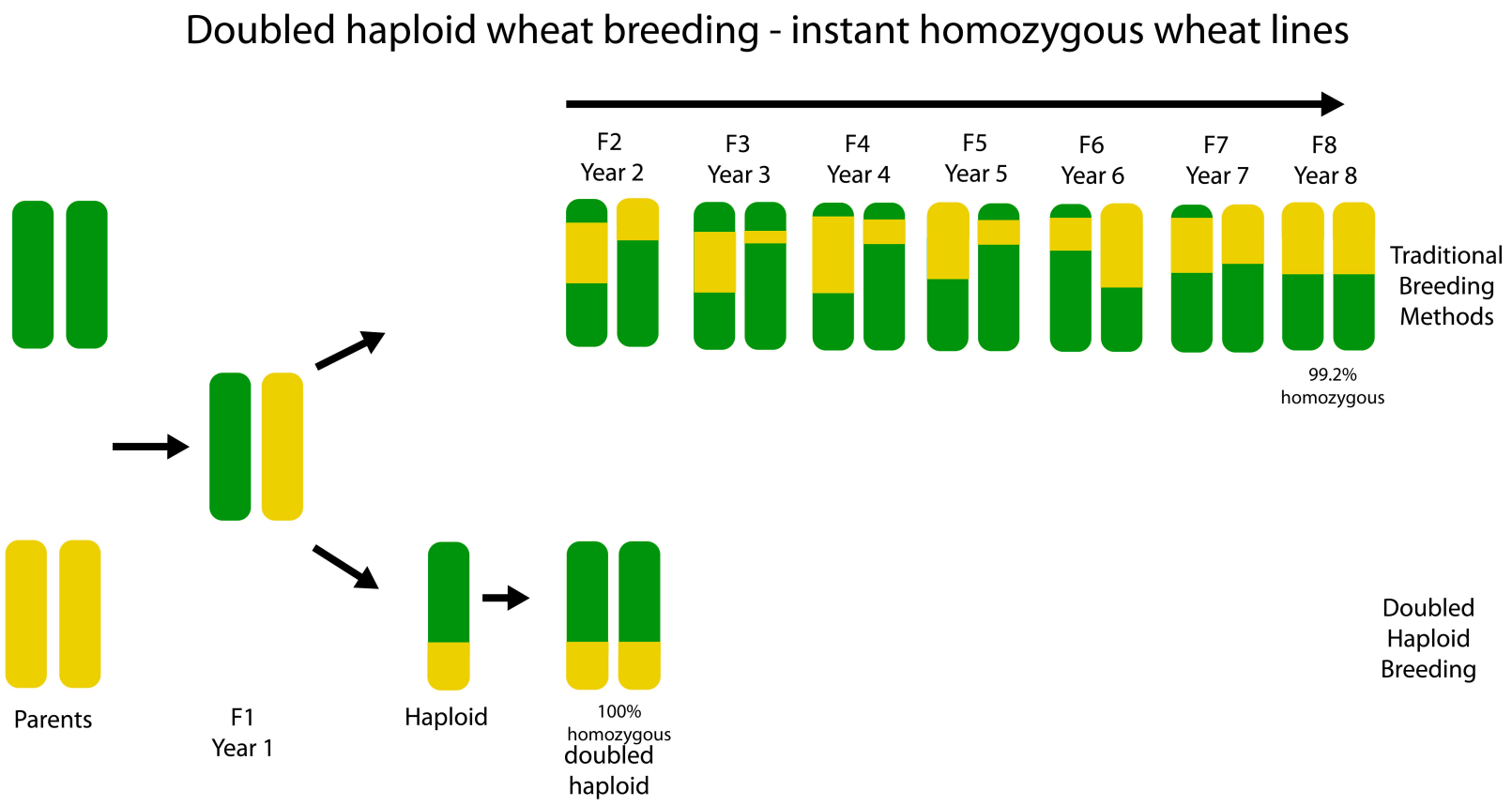
Crossing-over and recombination during meiosis

- Recombination fraction (θ) is the frequency with of a single recombinant event between two genes.

Mendel's 2nd Law

- Mendel's Second Law: the law of independent assortment - during meiosis, chromosomes assort randomly into gametes such that segregations of alleles of one gene is independent of alleles of another gene.
- As a consequence of this law, $E(\hat{\theta})$ will be 0.5 when two genes are located on different chromosomes or when they are widely separated on the same chromosome.
- When two genes are close together on the same chromosome, they do not assort independently and are said to be linked and $\theta < 0.5$.

Double Haploid population



	No. of markers	No. of lines
KUKRI × RAC875	6197	180
KUKRI × EXCALIBUR	5746	179
MACE × GLADIUS1	5054	207
RAC1548 × GLADIUS	5200	155
SCOUT × GLADIUS	5145	402
SCOUT × MACE	4950	255
AUS17840 × GLADIUS	5513	135
HALBERD × KENNEDY	6293	133
AUS17750 × GLADIUS	6160	125

Metrics

- A centimorgan (cM) is a genetic distance (as opposed to physical distance) that describes a recombination of 0.01.
- Kosambi mapping function is used for converting θ to cM which attempts to correct for multiple crossovers.
- The wheat genome is roughly 200cM per chromosome.

Metrics

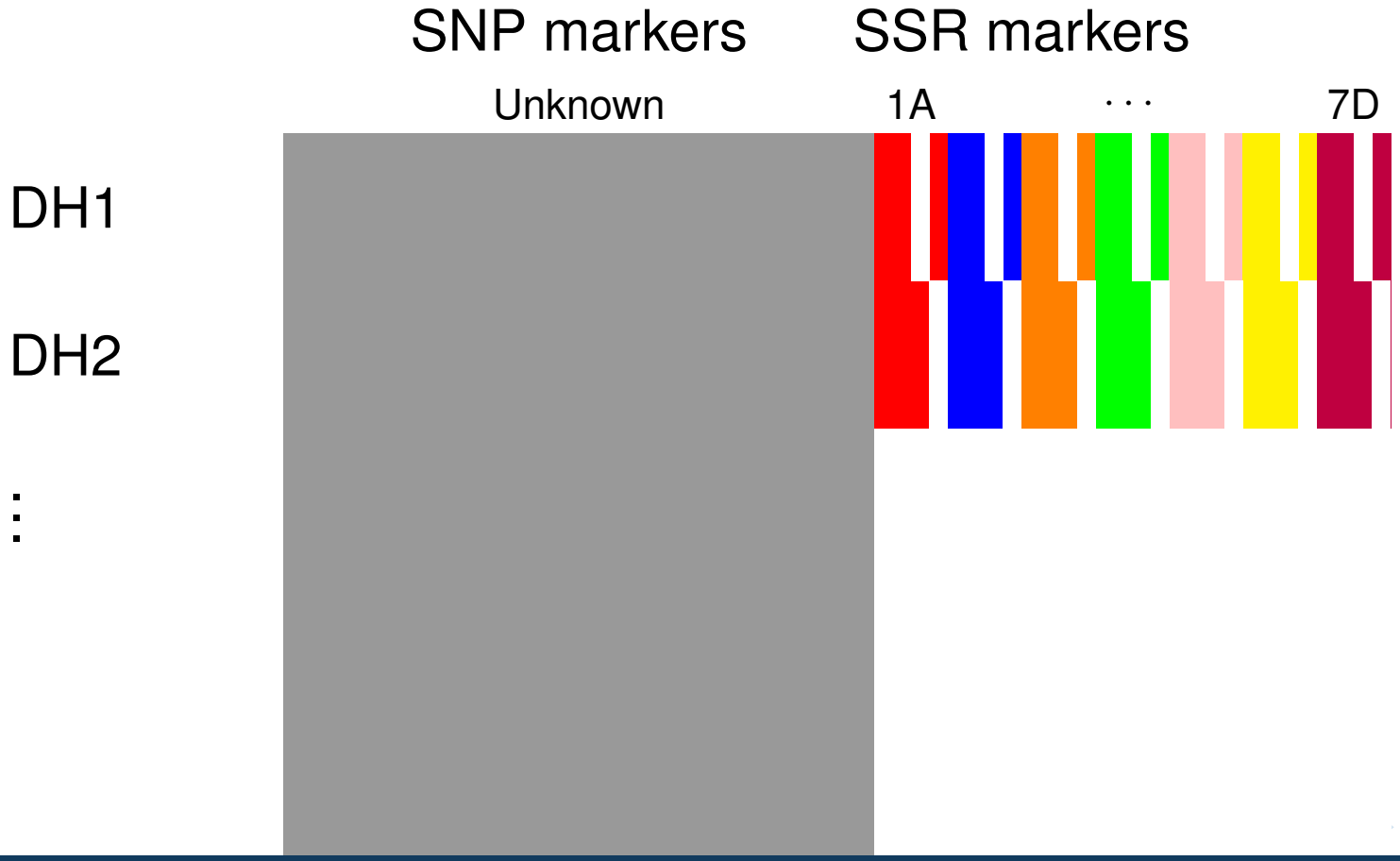
- The LOD score compares the likelihood of observed data if two loci are indeed linked to the likelihood of the same data purely by chance:

$$LOD = \log_{10} \frac{(1-\hat{\theta})^{NR} \times \hat{\theta}^{R}}{0.5^{NR+R}}$$

  where NR and R are the number of non-recombinant lines and R denotes the number of recombinant offspring.
- By convention LOD > 3.0 is considered evidence for linkage.

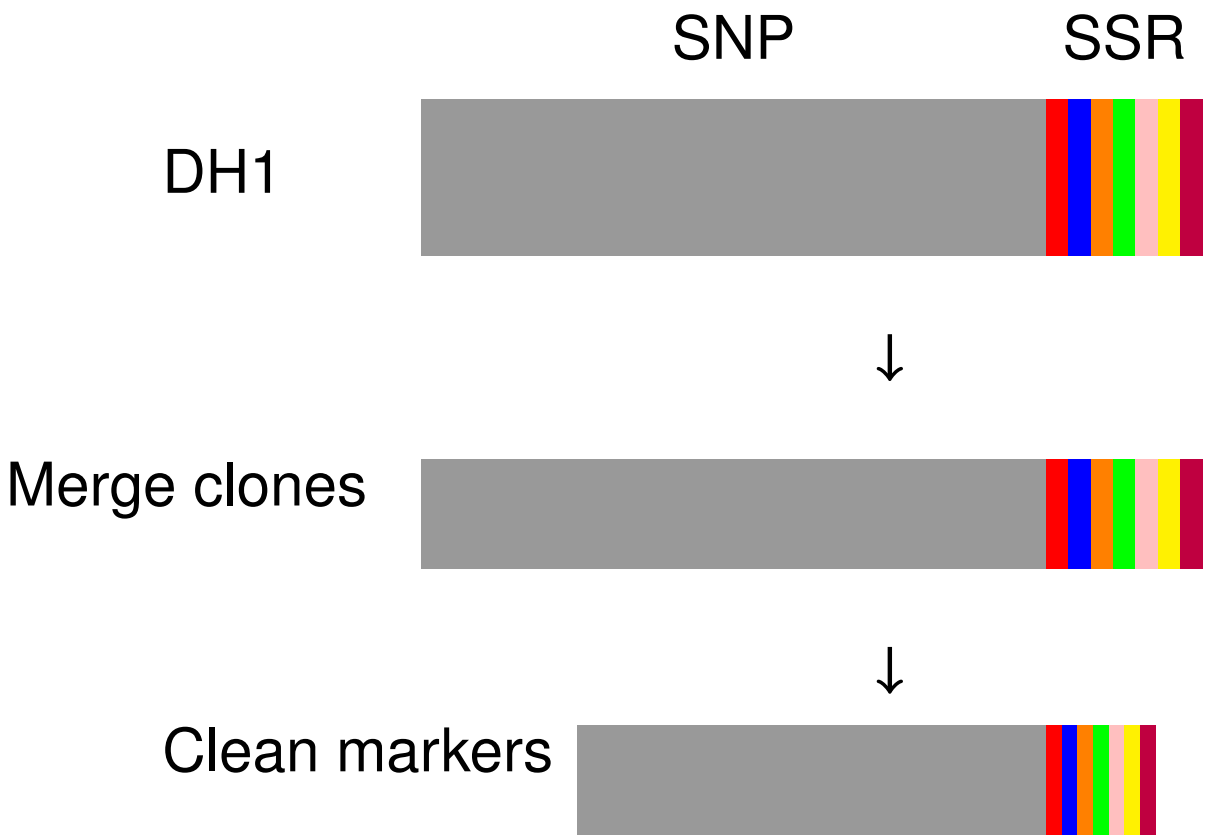
Data-Set

- We have 9 DH population with SNP markers for all.
- Two of these DH population also contain SSR markers with chromosome labels.



Data cleaning

- The following are done using R-package ASMap [Taylor and Butler, 2015].



Initial supervised learning

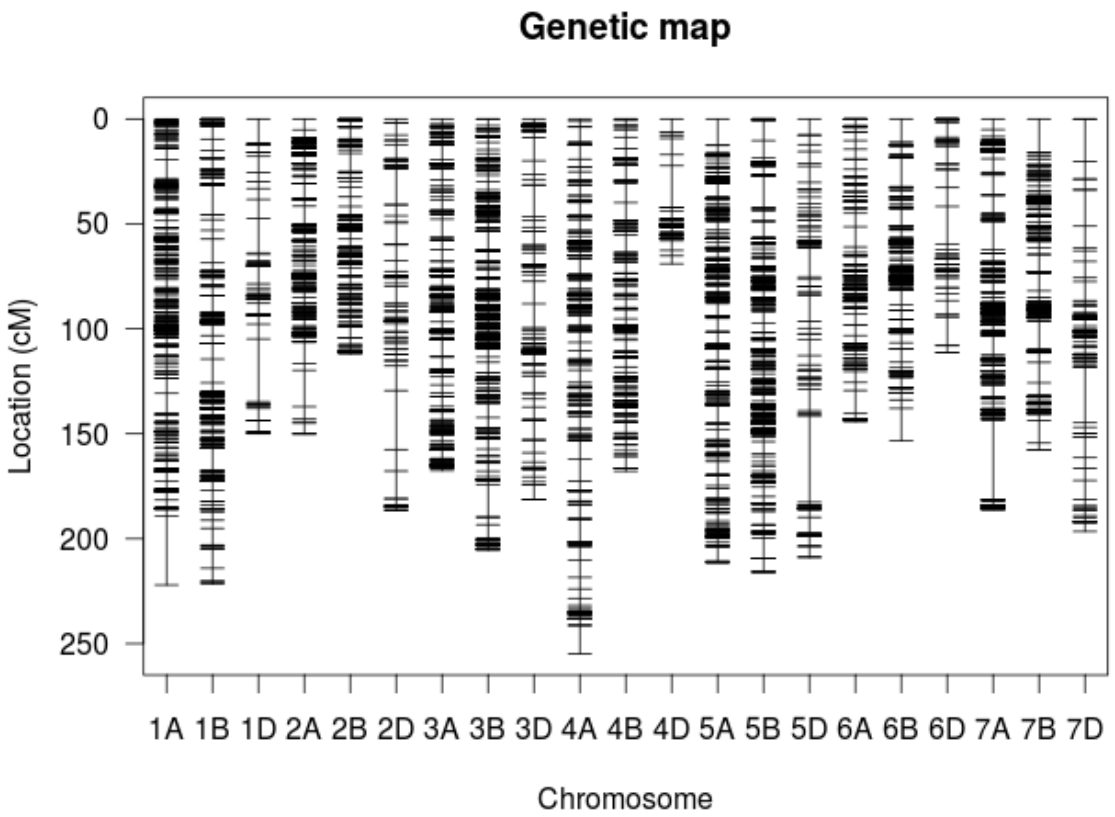
- The clustering of the “unanchored” SNP markers to “anchored” SSR markers (training set) are done under following assumptions:
  - The set of anchored markers have good coverage across the genome.
  - The labels of the anchored markers are correct.

Clustering chromosome groups

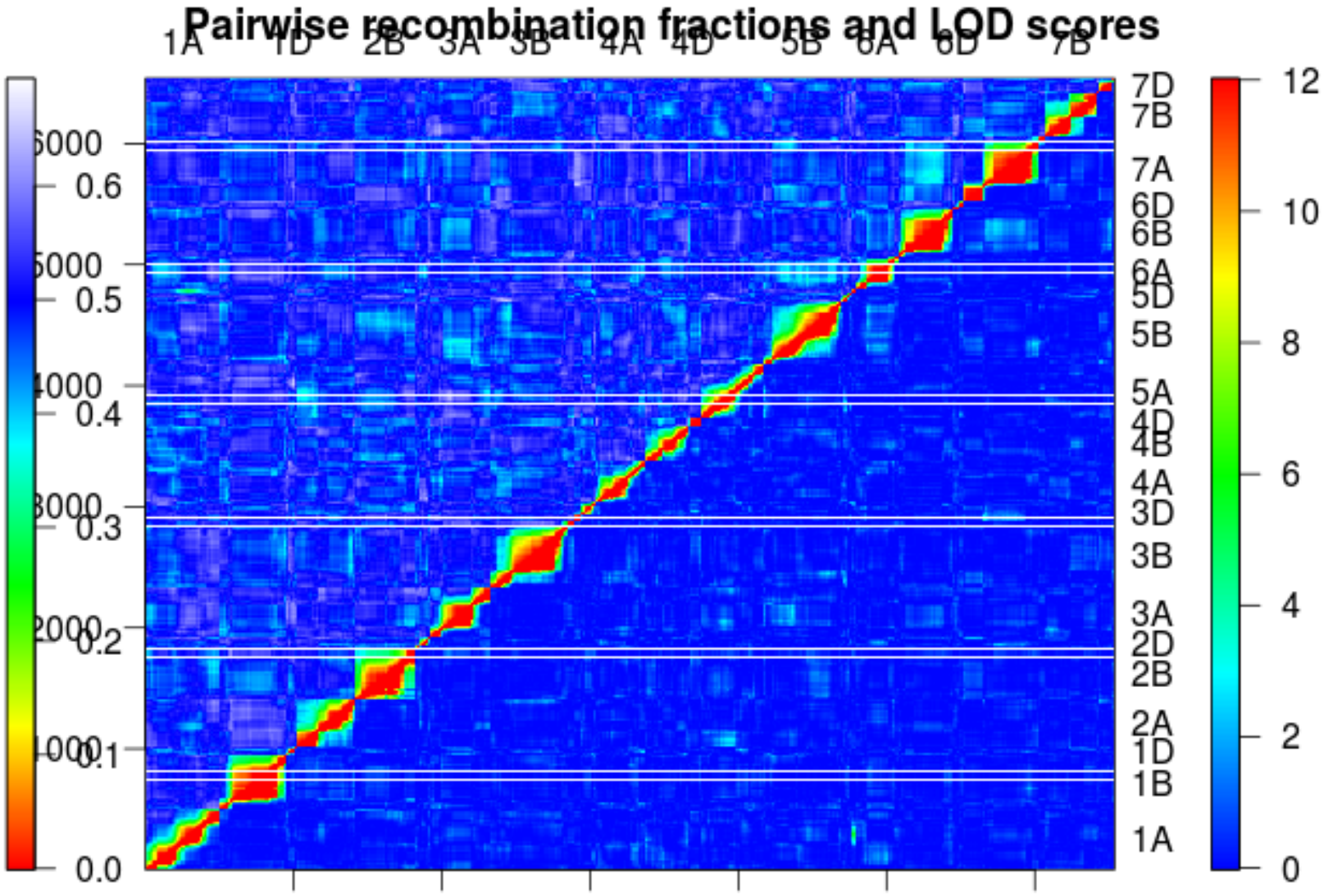
- The unanchored marker is assigned to the chromosome group of the anchor marker with minimum $\hat{\theta}$ out of all markers that have maximum $\hat{\theta}$ of 0.25 and minimum LOD of 3 (potentially linked markers).
- If there are no potentially linked markers in the anchored markers then the marker is linked to the unanchored marker with the least $\hat{\theta}$ out of potential linked markers in the unanchored group.

Linkage map

- The linkage map is constructed using ASMap that wraps the MSTMap algorithm [Wu et al., 2008a] in R keeping the chromosome labels from previous steps.



Linkage map


Pairwise recombination fractions and LOD scores
1A 1D 2B 3A 3B 4A 4D 5B 6A 6D 7B
7D 7B 7A 6D 6B 6A 5D 5B 5A 4D 4B 4A 3D 3B 3A 2D 2B 2A 1D 1B 1A
0.6 0.5 0.4 0.3 0.2 0.1 0.0
12 10 8 6 4 2 0

Supervised learning

- The process is repeated for DH2 independently of DH1.
- From the two linkage maps, the percentage of markers with the same chromosome assignments out of all common markers has a good concordance of 99.7% (3521/3530).
- These chromosome assignments are taken as new anchor markers to assign chromosomes for the other 7 DH populations.

Linkage Map Statistics

	No. of genotypes	No. of markers
KUKRI × RAC875	157	6538
KUKRI × EXCALIBUR	133	6123
MACE × GLADIUS1	176	5049
RAC1548 × GLADIUS	132	5175
SCOUT × GLADIUS	369	5145
SCOUT × MACE	226	4947
AUS17840 × GLADIUS	124	5510
HALBERD × KENNEDY	122	6292
AUS17750 × GLADIUS	116	6155

Supervised learning

- These 9 linkage maps, consisting of approx. 14K markers, are combined together into a consensus map using MergeMap [Wu et al., 2008b].
- The consensus map also provides a way to impute missing genomic data for QTL analysis.
- This map is also used for the QTL analysis to identify approximate locations of potential QTL.
- Furthermore the ordering of the map is exploited to estimate a correlation structure to the marker effects...

Supervised learning

- These 9 linkage maps, consisting of approx. 14K markers, are combined together into a consensus map using MergeMap [Wu et al., 2008b].
- The consensus map also provides a way to impute missing genomic data for QTL analysis.
- This map is also used for the QTL analysis to identify approximate locations of potential QTL.
- Furthermore the ordering of the map is exploited to estimate a correlation structure to the marker effects... to be continued ...

Acknowledgement

- Prof. Brian Cullis, Dr. Julian Taylor, Dr. Haydn Kuchel, and Adam Norman
- Australian Grain Technologies
- Australian Research Council for funding

Image courtesy of AGT

References I

Gianola, D., Perez-Enciso, M., and Toro, M. a. (2003).
On marker-assisted prediction of genetic value: Beyond the ridge.
Genetics, 163:347–365.

Morota, G., Boddhireddy, P., Vukasinovic, N., Gianola, D., and DeNise, S. (2014).
Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny tests for dairy cow health traits.
Frontiers in Genetics, 5(56).

References II

Smith, A. B. and Cullis, B. R. (2011).
Detecting QTL for photoperiod sensitivity in a doubled haploid Brassica napus population.
Technical report, SAGI Technical Report Series.

Taylor, J. and Butler, D. (2015).
ASMap: Linkage map construction using the MSTmap algorithm.

Verbyla, A. P., Cullis, B. R., and Thompson, R. (2007).
The analysis of QTL by simultaneous use of the full linkage map.
TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 116(1):95–111.

References III

Wu, Y., Bhat, P. R., Close, T. J., and Lonardi, S. (2008a).
Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph.
PLoS genetics, 4(10):e1000212.

Wu, Y., Close, T. J., and Lonardi, S. (2008b).
On the accurate construction of consensus genetic maps 1 1.
Computational Systems-Biology and Bioinformatics Conference, 7:285–296.

Yang, W. and Tempelman, R. J. (2012).
A bayesian antedependence model for whole genome prediction.
Genetics, 190(4):1491–1501.